5

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek pp. 32–38
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4514

A corpus linguistics perspective on language documentation, data, and the challenge of small corpora

Anke Lüdeling
Humboldt University of Berlin

This paper deals with issues of corpus design that might prove problematic for the study of under-resourced languages, e.g. in language documentation. It argues that it is not yet well understood which linguistic and extra-linguistic (predictor) variables cause linguistic variation (i.e. the response variable), which means that the scope of a linguistic finding cannot always be assessed. In order to deal with this problem, it is argued that we need a flexible corpus architecture with the option of adding meta-data to corpora/sub-corpora at any point in time.

1. INTRODUCTION. On an abstract technical level, there are no categorical differences between a large corpus for a well-researched language with many resources and a standardized orthography and a corpus of an endangered language or small variety without codified standards: In both cases one needs to represent a source text and annotations to it. Which levels of annotation are needed depends on the research question or purpose of the corpus; the corpus architecture needs to be flexible enough to represent annotation layers in different formats. A large number of corpora have been built in recent language documentation projects; and because many of these corpora are multi-modal and diverse, language documentation has had a significant impact on the development of corpus tools, formats, and architectures (see e.g. Auer et al. 2010, Wittenburg et al. 2010). Many of these tools, formats, and architectures are also used for other corpora – typically smaller corpora, corpora of spoken language, or corpora of non-standardized varieties, such as learner varieties, dialects, or historical language stages (see e.g. Ostler 2008 for an overview of smaller corpora, or Wittenburg 2008 for an overview of multimodal corpora) where linguistic analysis and annotation is crucial, and the methods developed for well-resourced languages or language-independent methods sometimes have fed back into language documentation projects (see e.g. Kirschenbaum et al., this volume, for a language-independent method of automatic morphological analysis). The same multi-layer standoff architecture and search tool (Annis, cf. Zeldes et al. 2009, Zipser & Romary 2010), for example, is used for a historical corpus of German (Petrova et al. 2009), deeply annotated corpora of African languages (Chiarcos et al. 2011), a modern German learner corpus (see below), or a large treebank of modern German newspaper texts.

On a more detailed level, however, there are issues of linguistic variation that can be understood only by working with many well-documented varieties of a language, and it might be that language documentation projects can learn from corpus work in the well-researched and richly-resourced languages.

2. CORPUS DESIGN. Research on linguistic variation is probably among the most interesting fields for which corpus data can be used, and I will focus on variation research here. It has been known for a long time that language production depends on many parameters – 'the same' (the variable) can be expressed in many different ways (the variants).[1] Research on variation focuses on the language-internal and language-external co-variables that influence the choice of one variant over another. In an ideal world, a corpus could be designed according to well-understood parameters that only depend on the research question at hand.[2] In addition to issues of time, money, and feasibility, this means that we would have to understand all the parameters that influence language production and could find portions of language for each necessary combination of parameters. This would mean that we select a sample (or several samples) of a language according to some design criteria and add these as meta-data to the corpus sources. I want to argue here that we are far from understanding which parameters influence variation and that we continue to discover new (not previously discussed) parameters that also influence linguistic variation. This means that we cannot work with a static set of meta-data categories.

Dedicated corpora as well as large reference corpora (such as 'national' corpora, the BNC, the ANC, DeReKo, etc.) that are not built for one specific research question include material for some of the possible combinations (such as n % of spoken informal dialogue or m % of written formal language, etc., see e.g. Hunston 2008 on corpus design). The rationale behind this is that for a given combination of parameters, each piece of material would be as good as any other piece of material. If we wanted to do research on, say, the use of adjectives in informal conversation we could take any informal conversation and generalize from that – there are countless publications on many aspects of the grammar of 'newspaper language', a given dialect, a given register, etc. that seem to (implicitly) follow this reasoning.[3]

However, linguistic research from many domains (sociolinguistics, dialectology, etc.) has discovered more and more parameters that seem to influence the choice of variants. These include among many others

- properties of the speaker: dialect, socio-economic factors, age, gender, education, etc., see e.g. Labov (1972, 2001)
- mode of communication: spoken, written, computer-mediated, etc., see e.g. the papers in Androutsopoulos & Beißwenger (2008) for CMC research or Siekmeyer (2011) for

[1] This is the view formulated in variationist sociolinguistics or in variationist models for language change (see e.g. Labov 1972, 1994, 2001 etc. or Rissanen 2008).

[2] In addition to corpora with a specific and documented design, we find opportunistic corpora and Web-based corpora where the composition is not known (such as the WaCky corpora, Baroni et al. 2009). While these might be used as 'example banks' or for NLP purposes, the lack of meta-information is problematic for more fine-grained variation research (see Lüdeling et al. 2007 for a discussion of the problems).

[3] This issue has sometimes been discussed under the header of 'representativeness'; see e.g. Biber (1993).

a corpus-based study of spoken versus written texts; mode is often discussed together with other parameters such as distance or level of formality (Koch & Oesterreicher 1994, Maas 2006)

- purpose of the communication: text type, functional variation, register, see e.g. Biber (1995, 2006).

There are several problems for corpus design here: First, for many of these co-varying variables it is unclear what the variants would be: Maas (2006), for example, distinguishes between three registers while Biber (1993, 1995, 2006) shows that many more distinctions can be made reliably. Second, without more research it is unclear which other factors play a role in language production. I want to illustrate this with an example from the German learner corpus Falko (Lüdeling et al. 2008, Reznicek et al. 2010). The corpus consists of written argumentative essays produced by advanced learners of German as a foreign language. It is well-known that the L1 has an influence on a learner's text production in the L2 (transfer, interference, see Ellis 2008). Golcher & Reznicek (2011, in preparation) show that a machine learning method that takes into account only substring frequencies is able to classify learner texts by L1 with high accuracy. In addition to using substring frequencies of the original texts, they also compare the part-of-speech annotation, again modeled as a string, and show that simply by looking at the frequencies of part-of-speech (POS) sequences, the L1 of a learner can be predicted with high accuracy. While this in itself might not be too surprising, Golcher & Reznicek use the same method to classify learner texts by their topic. All learners were asked to choose one of four topics; all other production parameters (time, setting, access to assistance, etc.) were kept equal for all texts. Figure 1 (from Golcher & Reznicek 2011: 31) shows that classification by topic is possible with an accuracy of around 80%, even by only taking into account POS sequences (thus not looking at lexical information). Automatic classification by topic is more reliable than classification by L1.

Findings such as these show that people use different grammatical means (represented by sequences of POS tags) when writing about different topics, even within the same text type (argumentative essay) written under the same circumstances.[4] This means that we might have to add 'topic' to the list of co-varying variables.[5] And if we look at further parameters, we might find further variables. The findings by Mosel (this volume), who shows that certain narrative styles are associated with specific stories in Teop, and Gullberg (this volume), who speaks about the different parameters that influence L2 acquisition, provide further cases in point.

3. CONSEQUENCES. Detailed studies that analyze variable after variable and their interaction with text production can only be done for languages with rich resources. But there

[4] The method here is a stylometric method developed in Golcher (2007). Golcher & Reznicek take care to eliminate any confounding L1 evidence that might have been introduced by the fact that learners with a given L1 tend to choose certain topics over others.

[5] Further studies are needed to find the reasons for these differences (Golcher & Reznicek, in preparation, provide some suggestions). It might, for instance, turn out to be the case that the differences are only indirectly due to the topic but point to register differences instead because some of the topics prompt the writers to argue more emotionally than others.

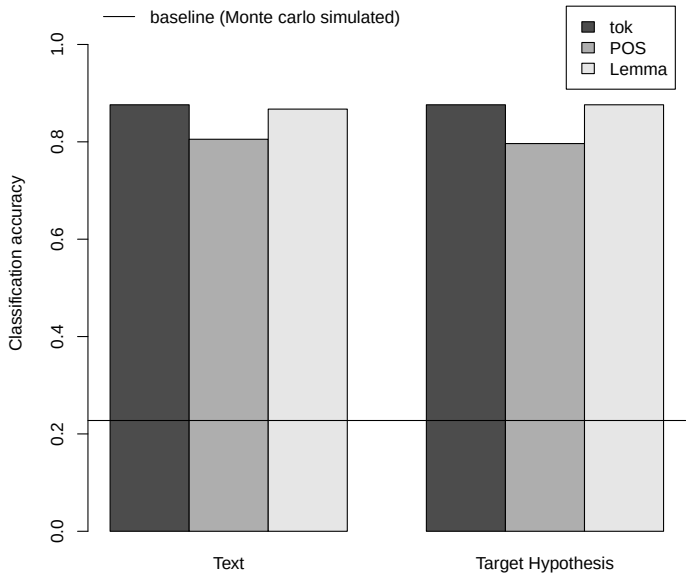

FIGURE 1: Classification of Falko texts by topic, using substring frequencies (figure taken from Golcher & Reznicek 2011: 31). The target hypothesis is a slightly modified version of the learner text where grammatical errors are corrected; results for classification are similar.

is no reason to assume that the findings from such research on richly-resourced languages will not in principle carry over to less-researched languages. Does this mean for language documentation that we only make grammatical statements about individual texts and do not generalize at all in such a context? This is clearly not the case because, as has been shown time and again. Speakers have a good awareness of appropriate variants (this is, of course, the basis for the fact that automatic classification works at all). But it is necessary that we understand which grammatical elements (on all grammatical levels, from sounds to text structure) are specific to external co-varying variables and which are not, and how variables interact with each other.[6]

On a technical level, we need powerful flexible corpus architectures where annotation layers and meta-data can be added at any point in time. Today very often a corpus can only be accessed via an interface and only the original corpus provider has the power to change it. If new linguistic analyses on this material are carried out and new co-varying variables

[6] There are many statistical models that can be used to reduce dimensions (such as principal component analysis) or can be used to look at the effect of different variables (also as random effects models). Interaction is crucial, and its effects have often been underestimated, cf. Gries, forthcoming.

are found after the corpus was first constructed, we need the option to add this information to the corpus. The technical means are there: Several multi-layer standoff architectures for corpora are being developed, and I am hopeful that the remaining technical issues (with respect to conversion frameworks and search tools, etc.) will be solved within the next few years. What seems to be just as necessary (but perhaps more difficult) is an awareness of the fact that we do not yet understand variation and cannot know all meta-data categories or annotation layers when we construct a corpus. This awareness would result in better corpus exchange and update functionalities and a fundamental work-flow change.

REFERENCES

ANC. American National Corpus. http://americannationalcorpus.org/ (11 December, 2011).

Androutsopoulos, Jannis & Michael Beißwenger. 2008. Introduction: Data and methods in computer-mediated discourse analysis. *Language@Internet* 5(2). http://www.languageatinternet.org/articles/2008/1609 (12 December, 2011).

Auer, Eric, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, Stefano Masnieri, Daniel Schneider & Sebastian Tschöpe. 2010. ELAN as flexible annotation framework for sound and image processing detectors. In *European Language Resources Association LREC 2010: Proceedings of the 7th International Language Resources and Evaluation*, 890–893. Paris: ELRA.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University Language: A corpus-based study of spoken and written register*. Amsterdam: John Benjamins.

BNC. British National Corpus. http://www.natcorp.ox.ac.uk/ (11 December, 2011).

Chiarcos, Christian, Ines Fiedler, Mira Grubic, Katharina Hartmann, Julia Ritz, Anne Schwarz, Amir Zeldes & Malte Zimmermann. 2011. Information structure in African languages: Corpora and tools. *Language Resources and Evaluation; Special Issue on African Language Technology* 45(3). 361–374.

DeReKo. Deutsches Referenzkorpus. http://www.ids-mannheim.de/kl/projekte/korpora/ (11 December, 2011).

Ellis, Rod. 2008. Language transfer. In Rod Ellis (ed.), *The Study of Second Language Acquisition*, 349–403. Oxford: Oxford University Press.

Falko. Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko (11 December, 2011).

Golcher, Felix. 2007. A new text statistical measure and its application to stylometry. In Matthew Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson (eds.), *Proceedings of the Corpus Linguistics Conference (CL2007)*, University of Birmingham. http://amor.cms.hu-berlin.de/~golcherf/cl07.pdf (12 December, 2011).

Golcher, Felix & Marc Reznicek. 2011. Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus. In Amir Zeldes & Anke Lüdeling (eds.), *QITL-4 – Proceedings of Quantitative Investigations in Theoretical Linguistics 4* QITL-4, 29–34. Berlin: Humboldt-Universität zu Berlin. http://edoc.hu-berlin.de/conferences/qitl-4/golcher-felix-29/PDF/golcher.pdf (12 December, 2011).

Golcher, Felix & Marc Reznicek. (in preparation). Stylometry and the interplay of L1 and text topic in second language texts.

Gries, Stefan Th. forthcoming. Statistische Modellierung. Zeitschrift für Germanistische Linguistik 2 (2012).

Gullberg, Marianne. this volume. Bilingual multimodality in language documentation data.

Hunston, Susan. 2008. Collection strategies and design decisions. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, vol. 1, 154–168. Berlin: Mouton de Gruyter.

Kirschenbaum, Amit, Peter Wittenburg & Gerhard Heyer. this volume. Unsupervised morphological analysis of small corpora: First experiments with Kilivila.

Koch, Peter & Wulf Oesterreicher. 1994. Schriftlichkeit und Sprache. In Hartmut Günther & Otto Ludwig (eds.), *Schrift und Schriftlichkeit / Writing and Its Use: Ein interdisziplinäres Handbuch internationaler Forschung / An Interdisciplinary Handbook of International Research*, 587–604. Berlin: Walter de Gruyter.

Labov, William. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1994. *Principles of Linguistic Change. Volume I: Internal Factors*. Oxford: Blackwell.

Labov, William. 2001. *Principles of Linguistic change. Volume II: Social Factors*. Oxford: Blackwell.

Lüdeling, Anke, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt & Maik Walter. 2008. Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 2(2008). 67–73.

Lüdeling, Anke, Stefan Evert & Marco Baroni. 2007. Using web data for linguistic purposes. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus Linguistics and the Web*, 7–24. Amsterdam: Rodopi.

Maas, Utz. 2006. Der Übergang von Oralität zu Skribalität in soziolinguistischer Perspektive / The Change from Oral to Written Communication from a Sociolinguistic Perspective. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier & Peter Trudgill (eds.), *Sociolinguistics. An International Handbook of the Science of Language and Society / Soziolinguistik. Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaf*, vol. 3, 2147–2170. Berlin: Walter de Gruyter.

Mosel, Ulrike. this volume. Creating educational materials in language documentation projects – creating innovative resources for linguistic research.

Ostler, Nicholas. 2008. Corpora of less studied languages. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, vol. 1, 457–483. Berlin: Walter de Gruyter.

Petrova, Svetlana, Michael Solf, Julia Ritz, Christian Chiarcos & Amir Zeldes. 2009. Building and using a richly annotated interlinear diachronic corpus: The case of Old High German Tatian. *Traitement Automatique des Langues* 50(2). 47–71.

Reznicek, Marc, Maik Walter, Karin Schmid, Anke Lüdeling, Hagen Hirschmann & Cedric Krummes. 2010. Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 1.0. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/pdf/Falko-Handbuch_Korpusaufbau%20und%20Annotationen.pdf (12 December, 2011).

Rissanen, Matti. 2008. Corpus linguistics and historical linguistics. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, vol. 1, 53–68. Berlin: Walter de Gruyter.

Siekmeyer, Anne. 2011. *Strukturelle Kategorien des sprachlichen Ausbaus bei Jugendlichen mit Deutsch als Erstsprache und Deutsch als Zweitsprache. Eine korpuslinguistische Untersuchung des Gebrauchs komplexer Nominalphrasen als Merkmale literater Strukturen in verschiedenen Registern*: Universität Saarbrücken dissertation.

Wittenburg, Peter. 2008. Preprocessing multimodal corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, 664–685. Berlin, New York: Mouton de Gruyter.

Wittenburg, Peter, Paul Trilsbeek & Przemek Lenkiewicz. 2010. Large multimedia archive for world languages. In *SSCS'10 - Proceedings of the 2010 ACM Workshop on Searching Spontaneous Conversational Speech, Co-located with ACM Multimedia 2010*, 53–56.

Zeldes, Amir, Julia Ritz, Anke Lüdeling & Christian Chiarcos. 2009. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009, Liverpool*, http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen/amir/pdf/CL2009_ANNIS_pre.pdf (07 June, 2012).

Zipser, Florian & Laurent Romary. 2010. A model-oriented approach to the mapping of annotation formats using standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*, Malta. http://hal.archives-ouvertes.fr/inria-00527799/en/ (12 December, 2011).

Anke Lüdeling
anke.luedeling@rz.hu-berlin.de